Queen Mary
University of London

Main Examination period 2022 – May/June – Semester B

MTH5120: Statistical Modelling I

You should attempt ALL questions. Marks available are shown next to the questions.

In completing this assessment:

- **You may use books and notes.**
- **You may use calculators and computers, but you must show your working for any calculations you do.**
- **You may use the Internet as a resource, but not to ask for the solution to an exam question or to copy any solution you find.**
- **You must not seek or obtain help from anyone else.**

All work should be **handwritten** and should **include your student number**.

The exam is available for a period of **24 hours**. Upon accessing the exam, you will have **2 hours** in which to complete and submit this assessment.

When you have finished:

- scan your work, convert it to a **single PDF file**, and submit this file using the tool below the link to the exam;
- e-mail a copy to **maths@qmul.ac.uk** with your student number and the module code in the subject line;
- with your e-mail, include a photograph of the first page of your work together with either yourself or your student ID card.

Please try to upload your work well before the end of the submission window, in case you experience computer problems. **Only one attempt is allowed – once you have submitted your work, it is final**.

IFoA exemptions. For actuarial students, this module counts towards IFoA actuarial exemptions. To be eligible for IFoA exemption, **your must submit your exam within the first 3 hours of the assessment period**.

Examiners: L. Shaheen, A. Zincenko

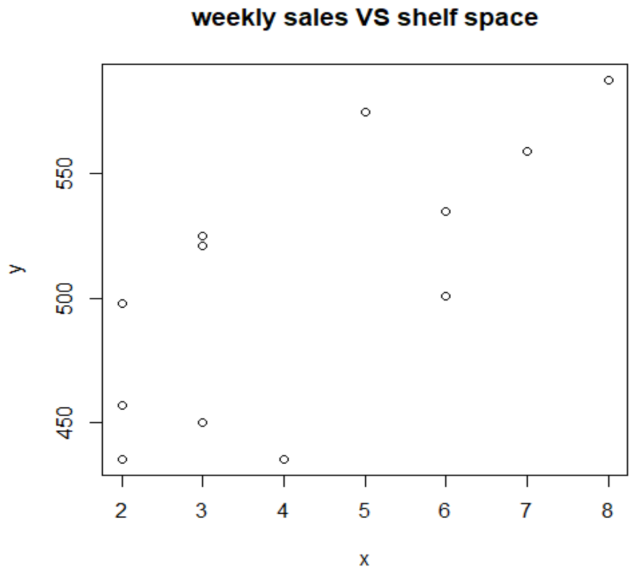
 Continue to next page

Question 1 [25 marks].
A baker is interested to find the relationship between the width of the shelf-space for her brand of cookies (x, in feet) and monthly sales (y) of the product in a supermarket. Hence, she fits a model relating monthly sales y to the amount of shelf space x her cookies receive that month. That is, she is fitting the model in the following way

$$y = \beta_0 + \beta_1 x + \epsilon$$

where $\epsilon \sim N(0, \sigma^2)$.

x (shelf space)	y (weekly sales)
3	535
2	425
6	575
5	639
3	450
8	630
4	435
2	498
6	534
3	530
2	457
7	559



Using the R, we obtained the following output.

```
> mody <- lm(y ~ x)
> summary(mody)

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-67.022 -31.346  -0.631  33.654  54.734

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  429.048     26.519  16.179 1.69e-08 ***
x             18.244      5.643   3.233  0.00898 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 39.2 on 10 degrees of freedom
Multiple R-squared:  0.511,     Adjusted R-squared:  0.4621
```

```
F-statistic: 10.45 on 1 and 10 DF,  p-value: 0.008979

> anova(mody)
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value   Pr(>F)
x          1  16059 16058.9  10.451 0.008979 **
Residuals 10  15366  1536.6
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Use the R output above to answer the questions below.

(a) By looking at the summary output, write down the fitted model. [**2**]

(b) Write down the formula to compute the 95% confidence interval for β_1? [**3**]

(c) Compute the 95% confidence interval for β_1. [**4**]

(d) Fill in the blanks in the following table.

Source of Variation	D F	Sum of Squares	Mean Square	F Value
Regression	1	SSR = 16059	MSR = 16058.9	?
Residual	12 - ? = ?	SSE = ?	MSE = $\frac{15366}{?}$ = ?	

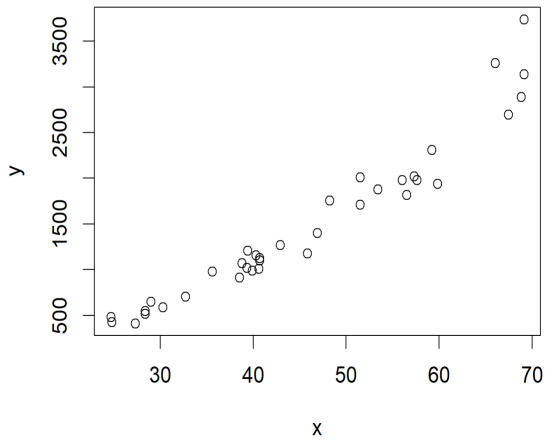
[**5**]

(e) (i) Write down the null hypothesis, that there is no effect on mean sales from increasing the amount of shelf space, versus a suitable alternative hypothesis. [**4**]

(ii) Compare the above value of F with the table value $F_{1,10,0.01}$. [**3**]

(iii) Comment on your findings. [**4**]

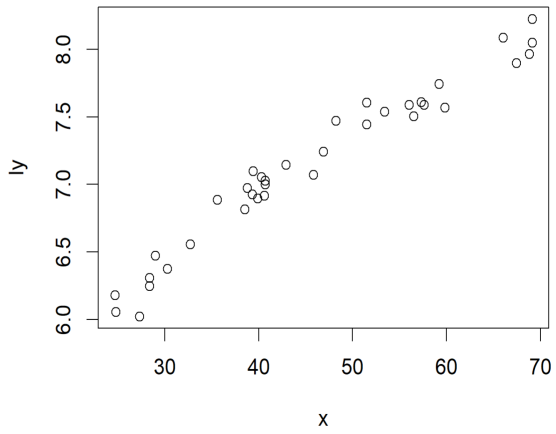
Question 2 [15 marks].
The thickness (x) and hardness (y) of 36 woods are plotted in the table below. We are interested in establishing the relationship between the y and x values. For these data, using R, we obtained the following output.

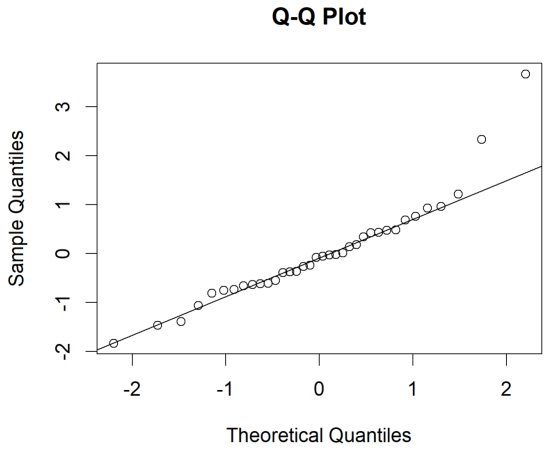


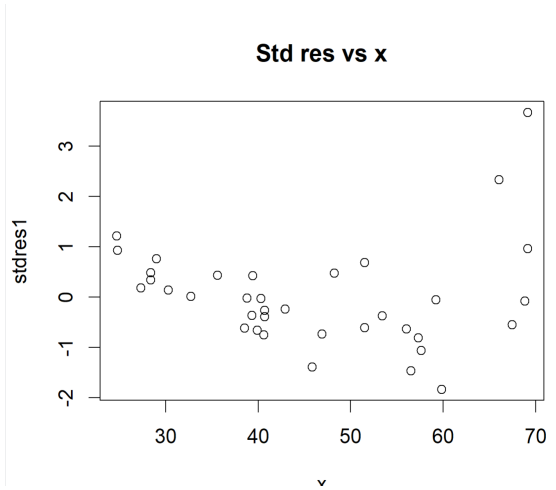

```
> mody1 <- lm(y ~ x)
> summary(mody1)
Call:
lm(formula = y ~ x)
Residuals:
    Min      1Q  Median      3Q     Max
-417.10 -142.03  -13.83  103.70  814.42
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1298.282    139.496  -9.307  7.1e-11 ***
x              61.127      2.927  20.882  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 235.2 on 34 degrees of freedom
Multiple R-squared:  0.9277,Adjusted R-squared:  0.9255
F-statistic:   436 on 1 and 34 DF,  p-value: < 2.2e-16

> anova(mody)
Analysis of Variance Table
Response: y
          Df   Sum Sq  Mean Sq F value    Pr(>F)
x          1 24117743 24117743  436.04 < 2.2e-16 ***
Residuals 34  1880575    55311
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> shapiro.test(stdres1)

           Shapiro-Wilk normality test

data:  stdres1
W = 0.9051, p-value = 0.004718
```

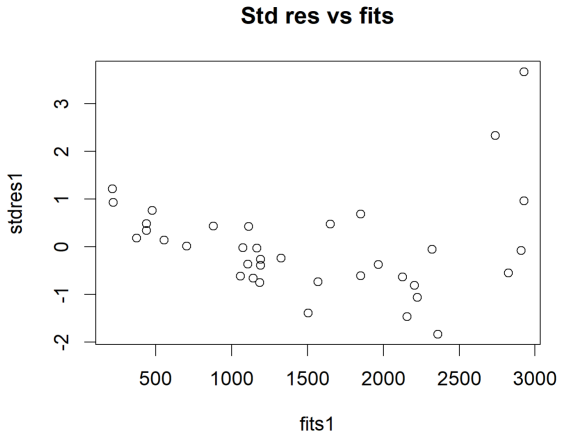
 Continue to next page

```
> ly <- log(y)
> plot(x,ly)
> stdres1 <- rstandard(mody1)
> fits1 <- fitted(mody1)
> plot(x,stdres1, main = "Std  res vs x")
> plot(fits1,stdres1, main = "Std res vs fits1")
> qqnorm(stdres1, main ="Q-Q plot")
> qqline(stdres1)
```









 Continue to next page

(a) Looking at the value of R^2 above, is this linear model a reasonable fit? [**3**]

(b) Viewing the residual plot, is there a possible problem with the constancy of variance? [**5**]

(c) Using the Q-Q plot and the Shapiro-Wilk test, check if there is a possible problem with the assumption of normality? [**4**]

(d) Looking at the plots above, is there any other transformation that you would like to consider? Give reasons for your answer. [**3**]

Question 3 [20 marks].

(a) Let $X_1, X_2, \cdots, X_n$ be random variables from a normal distribution with unknown mean μ and unknown variance σ^2. We are interested in finding the maximum likelihood estimates of μ and σ^2. Now let $\widehat{\mu}$ and $\widehat{\sigma^2}$ be the maximum likelihood estimates for μ and σ^2. The probability density function of X_i is given by

$$f(x_i; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x_i-\mu)^2}$$

for $-\infty < \mu < \infty$, $0 < \sigma^2 < \infty$ and $i = 1, 2, \cdots, n$.

Prove that

$$\widehat{\mu} = \frac{\sum_{i=1}^n x_i}{n} \text{ and } \widehat{\sigma^2} = \frac{\sum_{i=1}^n (x_i - \widehat{\mu})^2}{n}.$$

[**10**]

(b) Assume that $Y_1, Y_2, \cdots, Y_n$ are independent normal random variables with mean $\mu = \beta_0 + \beta_1 x_i$ and variance σ^2. The probability density function of Y_i is given by

$$g(y_i; \beta_0, \beta_1, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\{y_i-(\beta_0+\beta_1 x_i)\}^2}{2\sigma^2}}.$$

Prove that the values of the maximum likelihood estimates are

$$\widehat{\beta_0} = \overline{y} - \widehat{\beta_1}\overline{x},$$

$$\widehat{\beta_1} = \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^n (x_i - \overline{x})^2} = \frac{S_{xy}}{S_{xx}}$$

and

$$\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^n (y_i - (\widehat{\beta_0} + \widehat{\beta_1} x_i)^2.$$

[**10**]

Question 4 [25 marks].

A statistician is employed by a car industry. They wish to establish the relationship between a set of predictors and an outcome variable, which is mpg (miles per gallon) in this case. The original dataset is a data frame of 32 observations and 11 variables. Below is the snippet of the dataset.

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360.0	175	3.15	3.440	17.02	0	0	3	2
Valiant	18.1	6	225.0	105	2.76	3.460	20.22	1	0	3	1
Duster 360	14.3	8	360.0	245	3.21	3.570	15.84	0	0	3	4
Merc 240D	24.4	4	146.7	62	3.69	3.190	20.00	1	0	4	2
Merc 230	22.8	4	140.8	95	3.92	3.150	22.90	1	0	4	2
Merc 280	19.2	6	167.6	123	3.92	3.440	18.30	1	0	4	4
Merc 280C	17.8	6	167.6	123	3.92	3.440	18.90	1	0	4	4
Merc 450SE	16.4	8	275.8	180	3.07	4.070	17.40	0	0	3	3
Merc 450SL	17.3	8	275.8	180	3.07	3.730	17.60	0	0	3	3
Merc 450SLC	15.2	8	275.8	180	3.07	3.780	18.00	0	0	3	3

(a) The forward method for selecting the optimal set of predictors can use the AIC or BIC. Describe the model selection method, including the definition of the AIC and BIC. [4]

(b) By looking at the R output below, state which case corresponds to the AIC which to the BIC. Explain your reasoning. [5]

In the R output below, mpg is miles per gallon, wt is weight, cyl is number of cylinders, hp is gross horsepower.

```
Call:
lm(formula = mpg ~ wt + cyl + hp, data = mtcars)

Coefficients:
(Intercept)           wt          cyl           hp
38.75179     -3.16697     -0.94162     -0.01804

Call:
lm(formula = mpg ~ wt + cyl, data = mtcars)

Coefficients:
(Intercept)           wt          cyl
39.686       -3.191       -1.508
```

(c) Write down the model that corresponds to the R command below.

```
forward_aic <- step(lm(mpg ~ 1, data = mtcars), direction = "forward",
                    scope = formula(fit.full), k = 2, trace = 0)
```

[**2**]

(d) In the command above, what does k mean? [**2**]

Now, consider a new model where variables drat (real axel ratio), wt (weight), gear (number of forward gears), carb (number of carborators) are of interest. In the new model the response variable is y1 = 1/y, where y is mpg (miles per gallon).

A statistician employed a certain model selection procedure and has obtained the following R output.

```
Start:  AIC=-310.25
y1 ~ mtcars$drat + mtcars$wt + mtcars$gear + mtcars$carb

Df  Sum of Sq       RSS     AIC
- mtcars$drat  1 0.00000727 0.0014486 -312.09
- mtcars$gear  1 0.00004714 0.0014884 -311.23
<none>                      0.0014413 -310.25
- mtcars$carb  1 0.00026862 0.0017099 -306.79
- mtcars$wt    1 0.00088512 0.0023264 -296.93

Step:  AIC=-312.09
y1 ~ mtcars$wt + mtcars$gear + mtcars$carb

Df  Sum of Sq       RSS     AIC
- mtcars$gear  1 0.00007402 0.0015226 -312.50
<none>                      0.0014486 -312.09
- mtcars$carb  1 0.00027151 0.0017201 -308.60
- mtcars$wt    1 0.00108239 0.0025309 -296.24

Step:  AIC=-312.5
y1 ~ mtcars$wt + mtcars$carb

Df Sum of Sq       RSS     AIC
<none>                      0.0015226 -312.50
- mtcars$carb  1 0.0002176 0.0017402 -310.22
- mtcars$wt    1 0.0045233 0.0060458 -270.37
```

(e) By looking at the output above, state which procedure the statistician has employed. Describe the procedure. State the reason as to why the procedure stops at the point that it does. [**6**]

(f) What is multicollinearity? Why is multicollinearity problematic? [**4**]

(g) We have calculated vif (variance inflation factor). State what conclusion we can draw with respect to collinearity of predictors in the model. [**2**]

```
> reduced.modely1

Call:
lm(formula = y1 ~ mtcars$wt + mtcars$carb)

Coefficients:
(Intercept)    mtcars$wt  mtcars$carb
0.005186     0.013657     0.001815


> vif(reduced.modely1)
x2       x4
1.223761 1.223761
```

Question 5 [15 marks].

Below, we see snippet of the dataset that consists of 128 observations and 15 variables. We present two models describing the linear relationship between the BodyFat and other variables.

	Density	BodyFat	Age	Weight	Height	Neck	Chest	Abdomen	Hip	Thigh	Knee	Ankle
1	1.0708	12.3	23	154.25	67.75	36.2	93.1	85.2	94.5	59.0	37.3	21.9
2	1.0853	6.1	22	173.25	72.25	38.5	93.6	83.0	98.7	58.7	37.3	23.4
3	1.0414	25.3	22	154.00	66.25	34.0	95.8	87.9	99.2	59.6	38.9	24.0
4	1.0751	10.4	26	184.75	72.25	37.4	101.8	86.4	101.2	60.1	37.3	22.8
5	1.0340	28.7	24	184.25	71.25	34.4	97.3	100.0	101.9	63.2	42.2	24.0
6	1.0502	20.9	24	210.25	74.75	39.0	104.5	94.4	107.8	66.0	42.0	25.6

	Biceps	Forearm	Wrist
1	32.0	27.4	17.1
2	30.5	28.9	18.2
3	28.8	25.2	16.6
4	32.4	29.4	18.2
5	32.2	27.7	17.7
6	35.7	30.6	18.8

(a) By looking at the `R` output, state whether one should include extra parameters in the model. [**2**]

```
Analysis of Variance Table

Model 1: x$BodyFat ~ x$Neck
Model 2: x$BodyFat ~ x$Neck + x$Weight + x$Height
Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1    250 13348.1
2    248  9461.4  2    3886.7 50.939 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(b) Let us consider just Hip, Forearm and Wrist as our predictors. How many possible linear models that predict BodyFat can one build? [**2**]

(c) Consider the two models

$$Salary = 6366 + 9.3\,Age - 329.56\,Male, \quad R^2 = 1.29, \quad \hat{\sigma}^2 = 1099 \qquad (1)$$

and

$$\log(Salary) = 5.342 + 0.012\,Age - 0.321\,Male, \quad R^2 = 0.162, \quad \hat{\sigma}^2 = 1.231. \quad (2)$$

(i) Interpret the coefficient for Male in each model. [**4**]

(ii) Would it be correct to say that the second model is preferred over the first? Explain your reasoning. [**3**]

Consider another model

$$\log(Salary) = 3.54 + 0.127\,Age - 0.321\,Male, \quad R^2 = 0.280, \quad \hat{\sigma}^2 = 0.757. \quad (3)$$

(iii) Is model (3) better than model (2)? Why? [**4**]

End of Paper.